卷　　　　分析化学　　　　期
年　月　　Chinese Journal of Analytical Chemistry　　起止页码

DOI:**********

基于近红外光谱和特征波长选择的香榧陈籽快速无损判别

范郑欣，昝佳君，杜　园，孙通*

（浙江农林大学光机电工程学院　杭州　311300）

摘　要　目的：香榧陈籽由于储存期间不饱和脂肪酸氧化，导致其口感降低，品质变差。不法商人为谋取暴利，将香榧陈籽掺入新籽售卖，侵害消费者利益，需要一种快速无损鉴别方法。方法：本研究利用近红外光谱技术对香榧陈籽进行快速无损判别研究。采用两种近红外光谱仪在200-1160nm及900-1700nm波段范围内采集带壳香榧样本的光谱，使用9种方法对光谱数据进行预处理，然后利用区间优化选择算法（ICO）、竞争自适应重加权采样（CARS）、连续投影算法（SPA）和变量组合种群分析(VCPA)四种波长选择方法筛选香榧陈籽的光谱特征变量，并应用线性判别分析（LDA）、支持向量机（SVM）、反向神经网络（BP）方法建立香榧陈籽的判别模型。结果：对于光谱仪1，CARS方法为最优的波长选择方法，CARS-SVM模型的性能最优，其预测集的敏感性、特异性和准确率均为100%。对于光谱仪2，标准化和SNV为较优的预处理方法，VCPA变量选择方法优于其它三种方法，所建立的最优模型为VCPA-BP，其模型的预测集敏感性、特异性和准确率分别为98.18%、93.02%、和95.04%。结论：由此可知，两光谱仪数据建立的判别模型均能较好地实现香榧陈籽的判别，光谱仪1的模型总体上性能优于光谱仪2。本研究可为香榧陈籽快速无损判别提供一种检测方法，有效保障香榧的品质。

关键词　近红外光谱(near-infrared spectroscopy)；香榧籽(Torreya grandis seeds)；陈籽(Toreeya stale seeds)；波长变量选择(Wavelength variable selection)；判别模型(Discrimination model)。

香榧为红豆杉科榧树属，是中国稀有木本经济树种，主要生长在中国南方海拔1400米以上的湿润地区，适应温暖多雨的黄壤、红壤等土壤，广泛分布于浙江、湖南、福建、贵州等地[1]。香榧籽为香榧果实，其营养丰富，含有多种脂肪酸、氨基酸和矿物元素，尤其富含钾元素，对降低胆固醇、促进脂溶性维生素吸收、改善肠道功能有益[2]。香榧陈籽为上一年采摘的香榧籽，由于香榧籽含油量高，存储一年后将会发生不同程度的氧化导致不饱和脂肪酸减少，会影响香榧籽品质和口感，其售价也会大大降低。不法商人为谋取利益，

****-**-**收稿；　**-**-**接受
浙江省属高校基本科研业务费专项资金资助(No. 2021TD002)、浙江省重点研发(No. 2020C02019)。
E-mail: suntong980@163.com

将香榧陈籽掺入正常香榧籽(新籽)进行售卖，侵害消费者的利益。因此，非常有必要对香榧陈籽进行快速无损检测。

近红外光谱技术作为一种快速、高效、便捷的无损检测手段，近年来已在农产品[3,4]、食品[5,6]、药品[7,8]、石油化工[9,10]等领域开展了较多的研究。对于坚果内部品质及缺陷检测，国内外研究者利用近红外光谱技术也进行了一定的研究。Jiaqi Hu 等[11]利用近红外光谱技术建立了新鲜板栗水分和水溶性糖的 PLS 定量检测模型，校正集和验证集的 R^2 均大于 0.9，均方根误差低于 0.05。刘洁等[12]建立了带壳板栗蛋白质含量的近红外模型，模型的校正集及验证集的相关系数分别为 0.8270 和 0.7655,均方根误差分别为 2.27%和 2.35%。张严等[13]利用近红外技术检测花生籽粒的脂肪酸含量，由此建立了 5 种脂肪酸的定量预测模型，其中油酸和亚油酸的定标模型性能较好,外部验证决定系数分别为 0.9401 和 0.9487,RPD 值均大于 2.5。棕榈酸、花生酸和山嵛酸定标模型的外部验证决定系数均大于 0.8。Giovanna Canneddu 等[14]通过傅立叶变换近红外光谱和 PCA-LDA 方法预测澳洲坚果过氧化值和酸度指数，其模型的预测集决定系数和 SEP 分别为 0.72 和 3.45。澳洲坚果分类模型的准确率达到 93.2%。Roberto Moscetti 等[15]采用近红外光谱技术结合 LDA 和 QDA 算法，成功建立了板栗霉变的识别模型。模型分类错误率低，假阴性、假阳性和总错误率分别为 2.42%、2.34%和 2.38%。Lívia C. Carvalho 等[16]使用主成分分析-线性判别分析（PCA-LDA）和遗传算法-线性判别分析（GA-LDA）建立澳洲坚果品种的分类模型。带壳澳洲坚果 PCA-LDA 分类模型的准确率为 93.2%，GA-LDA 分类模型的准确度高于 94.44％。Satyabrata Ghosh 等[17]采用近红外光谱和 PLSDA 方法获成功地将花生和松子与其他坚果（杏仁）和谷物（芝麻和亚麻籽）区分开来。由上述研究可知，近红外光谱技术对坚果内部品质与缺陷检测的研究主要集中在板栗、花生等薄壳坚果或果仁上，而对香榧等厚壳坚果的检测研究较少，目前尚未有香榧陈籽的近红外检测研究，亟需深入探索。

本研究利用近红外光谱技术香榧陈籽进行快速无损检测研究。采用两种不同波段的近红外光谱仪采集香榧样本光谱，利用 9 种预处理方法对光谱进行预处理，然后筛选香榧样本的特征光谱，最后将两光谱仪的信息应用线性判别分析、支持向量机及 BP 神经网络方法建立香榧陈籽的判别模型。

1 材料与方法

1.1 试验样本

香榧新陈籽样本均采购于浙江省绍兴赵家镇会稽山香榧群的核心区，为炒制加工处理后的香榧籽。随机选择大小均匀、外观正常的 2022 年香榧陈籽样本 216 个及 2023 年香榧新籽样本 216 个，将试验样本依次标号并保存于 4℃冷库中直到实验使用。在光谱采集之前，将香榧样本从冷库中取出并在室温环境中放置 12 小

时，以避免温度等因素对实验的干扰。

将上述样本采用 kennard-stone 法按照 2:1 划分为校正集和预测集，校正集样本用于建立判别模型，预测集样本用于验证判别模型。经样本划分后，校正集共 281 个样本，其中 130 个新籽，151 个陈籽。预测集共 151 个样本，86 个新籽，65 个陈籽。

1.2 试验装置与光谱采集

试验所用的香榧籽近红外透射检测装置如图 1 所示，采用两种光谱仪对香榧籽进行检测。光谱仪 1 为荷兰 Avantes 公司的 AvaSpec-HS1024×122TEC 型超高灵敏度光纤光谱仪，仪器波长范围为 200-1160 nm，探测器类型为热电制冷、薄型背照式 CCD，信噪比 1000：1。光谱仪 2 为上海复享光学股份有限公司的 NIR17S 型近红外光谱仪，波长范围为 900-1700 nm，探测器类型为热电制冷高速线列 InGaAs CCD，信噪比 3000：1。光源为卤素灯，功率为 150W。光纤为 Avantes 公司的 FC-UV600-2-ME 型光纤，芯径为 600μm。

光谱采集时,打开近红外光谱仪预热 30 min 以保证所测光谱的稳定性。然后将香榧样本放置在载物台上，采用光谱仪以透射方式采集一次光谱，顺时针旋转 90°再采集 1 次光谱，取两次光谱的平均值作为该样本的光谱。两种光谱仪均按照上述方式对香榧籽样本进行光谱采集，AvaSpec-HS 光谱仪的积分时间为 50 ms，NIR17S 光谱仪的积分时间为 100 ms，两光谱仪的扫描次数和平滑点数均为 1 次和 1 点，光谱参比为直径 30 mm 的聚四氟乙烯球。为方便叙述，后续中 AvaSpec-HS 光谱仪简称为光谱仪 1，NIR17S 光谱仪简称为光谱仪 2。

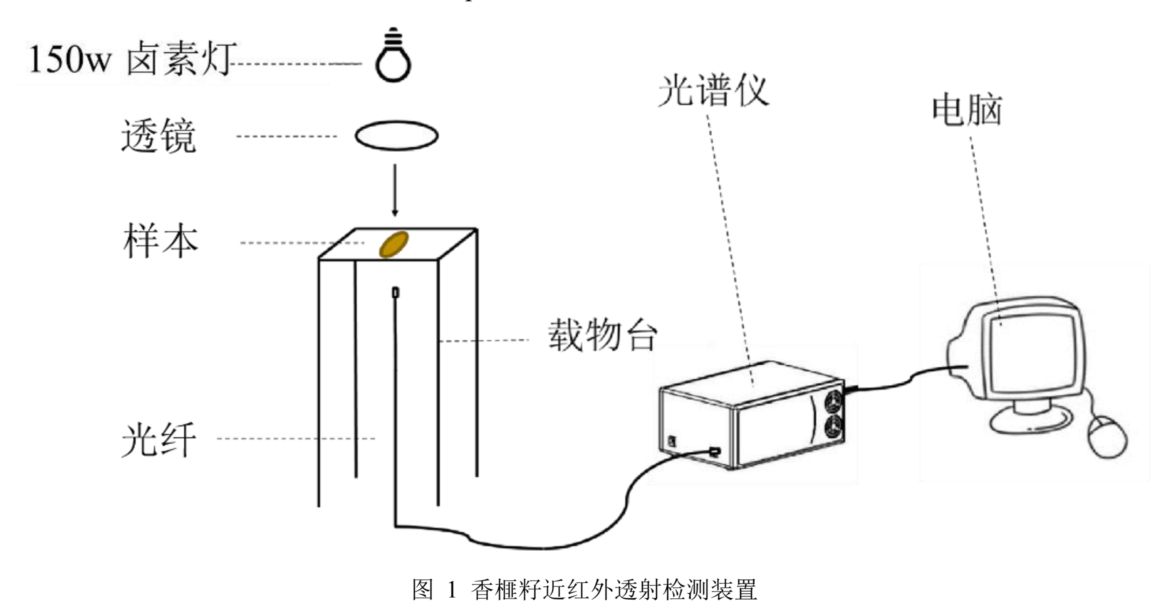


图 1 香榧籽近红外透射检测装置

Fig.1 Near infrared transmission detection device of Torreya grandis seeds.

1.3 数据处理与分析

1.3.1 光谱预处理

原始光谱数据中除了携带样本自身的化学信息外，还掺杂了其他噪声和无关信息，不同的预处理方法可

以消除光谱中的这些干扰因素，以此确保模型的稳定性和准确性[18,19]。本研究中，分别采用高斯滤波(Gaussian filter,GF)、中值滤波(Median Filter,MF)、SG(Savitzky-Golay Smoothing,SG)平滑、归一化(Normalize,Nor)、一阶导数(first derivative,FD)、多元散射校正(Multivariate scattering correction,MSC)、变量标准化(standard normalized variate,SNV)、去趋势(De-trending,De)和标准化(standard) 9 种方法对两种光谱仪采集的香榧籽原始光谱进行预处理，然后应用 LDA、SVM 及 BP 神经网络方法建立香榧陈籽的判别模型，并根据判别模型结果以确定较优的预处理方法。

1.3.2 特征波长选择方法

竞争自适应重加权采样(competitive adaptive reweighted sampling，CARS)方法[20]是一种基于进化论的"适者生存"原则进行特征变量筛选的方法。CARS 算法的基本原理是通过自适应重加权采样技术(ARS)筛选 PLS 模型中回归系数绝对值大的 波长点，去掉权重小的波长点，利用交互验证选出分类器中代价敏感错误率(cost-sensitive)最低的子集，可有效寻出最优变量组合。

区间组合优化算法(interval combination optimization，ICO)[21]是一种在模型集群分析框架下的新型优化区间选择算法。该方法将样本的近红外光谱划分为 N 个等宽区间，通过加权自举采样(Weighted Bootstrap Sampling, WBS)结合模型集群分析生成不同波长区间组合的子集。这些子集的初始权重设为 1，并使用偏最小二乘判别分析(PLS-DA)与 10 折交叉验证计算各区间的误差。基于误差最小的准则，从中提取一定比例的子集，赋予权重后迭代进行计算，直至标准差开始上升。每次迭代过程中，应用局部搜索策略优化所选波长区间的宽度。最终，误差值最小的波长区间组合被视为最佳波长区间。

连续投影算法(successive projections algorithm，SPA)[22]是一种基于前向变量选择并从光谱矩阵中选择共线性最小变量组合的算法。通过最小化矢量空间共线性，采用向量投影分析挑选最大向量，最后经过模型校正筛选光谱的特征变量，减少了模型的冗余度，提高了模型的稳定性和可靠性。

变量组合集群分析(variable combination population analysis，VCPA)[23]是一种新兴的波长变量选择方法，其充分考虑了变量集之间可能存在的影响。该方法由两个关键因素组成，一是指数递减函数(EDF)，该函数主要用于确定变量的数量，以保持在迭代过程中不断缩小变量空间。二是二元矩阵抽样(Binary matrix sampling,BMS)策略，每一代 EDF 的运行中，采用 BMS 策略为每个变量提供相同的选择机会，同时生成不同变量组合的子集用以构建子模型，并根据误差最小原则确定较优的变量子集。

本试验中，CARS 算法的主成分个数设为 8，采用 10 折交叉验证，迭代次数为 50 次。SPA 算法中，选择的特征波长数变量数范围设为 8-45 个。对于 ICO 算法，主成分数设为 10 个。对于 VCPA 算法，主成分数设为 10 个，采用 10 折交叉验证，EDF 迭代次数为 50，BMS 运行次数为 1000 次。

1.3.3 建模方法

LDA 是一种常见的识别算法，它在特征空间中找到一个投影轴，使得同类样本的投影点尽可能地接近，不同类别之间的投影点尽可能地远离[24,25]。SVM 是一种用于识别和回归分析的监督学习模型。其基本原理是通过找到一条最佳的分割超平面来最大化两类数据点之间的间隔，同时使得距离分割超平面最近的数据点(支持向量)的距离最大化[26,27]。BP 神经网络(Back Propagation)是一种按照误差逆向传播算法训练的多层前馈神经网络模型。由输入层、隐藏层和输出层三部分组成。其结构特点是层与层之间的连接均为全连接形成复杂的网络权重矩阵[28,29]。

本研究中，将香榧陈籽的类别值记为 1，香榧新籽的类别值记为 0，然后采用 LDA、SVM 及 BP 神经网络方法建立香榧陈籽的判别模型。

1.3.4 模型评价

为了评估所建立的判别模型性能，分别采用校正集和预测集的敏感性、特异性和准确率对判别模型性能进行评价。模型性能评价指标定义如式(1)-(3)。

敏感性（%）=$TP/(TP+FN)\times 100$ （1）

特异性（%）=$TN/(TN+FP)\times 100$ （2）

准确率（%）=$(TP+TN)/(TP+TN+FP+FN)\times 100$ （3）

式中：TP、TN、FP 和 FN 分别表示真阳性、真阴性、假阳性和假阴性。本研究中规定阳性为香榧陈籽，阴性为香榧新籽，即敏感性代表模型辨识香榧陈籽的能力，特异性代表模式辨识香榧新籽的能力。通常，敏感性、特异性及准确率越高，说明该方法对样本的识别和分类能力越强。

2 结果与讨论

2.1 近红外光谱分析

由于香榧籽的近红外光谱在 200-645 nm 及 1600-1700 nm 波长范围的噪声比较大，因此选用 646-1100 nm 及 900-1600 nm 波长范围的香榧籽近红外光谱用于分析。图 2 为两种光谱仪采集的香榧新陈籽平均近红外光谱。由图 2(a)可知，对于光谱仪 1，香榧新籽与香榧陈籽的平均光谱在 646-1100 nm 波长范围内变化趋势基本相似，但香榧新籽的吸光度小于陈籽；吸收峰所在的波长位置基本相同，主要为 900-945 nm 范围内的明显吸收峰及 730-750 nm 与 980-1050 nm 范围的小吸收峰。730-750 nm 附近的弱吸收峰为甲基 C-H 键的四级倍频；对于 900-945 nm 范围的吸收峰，913 nm 附近为甲基 C-H 键三级倍频，934 nm 附近为亚甲基 C-H 键级倍频；1000 nm 左右区域为醇类物质的 O-H 键伸缩振动的二级倍频。

由图 2(B)可知，对于光谱仪 2，香榧新籽与香榧陈籽的平均光谱在 900-1600 nm 波长范围内变化趋势基本相似，大致呈“W”型。香榧陈籽与新籽的吸光度存在交叉重叠，在 1050-1130 nm 及 1270-1320 nm 波段范围，香榧陈籽吸光度低于香榧新籽，而在 1190-1220 nm 及 1400-1600 nm 波段范围，香榧陈籽吸光度高于香榧新籽。光谱在 920 nm 及 934 nm 附近存在较小的吸收峰，920 nm 吸收峰归因于甲基 C-H 键的三级倍频吸收，934 nm 为亚甲基 C-H 键的吸收峰。此外，光谱在 1170 nm 和 1220 nm 区域也展现出明显的吸收峰，为甲基与亚甲基中 C-H 双键的二级倍频吸收谱带。在 1440-1460 nm 区间出现了属于 N-H 键反对称伸缩的一级倍频吸收，具体位于 1446 nm 附近。1440 nm 处吸收峰为水分子中 O-H 键伸缩振动的一级倍频。

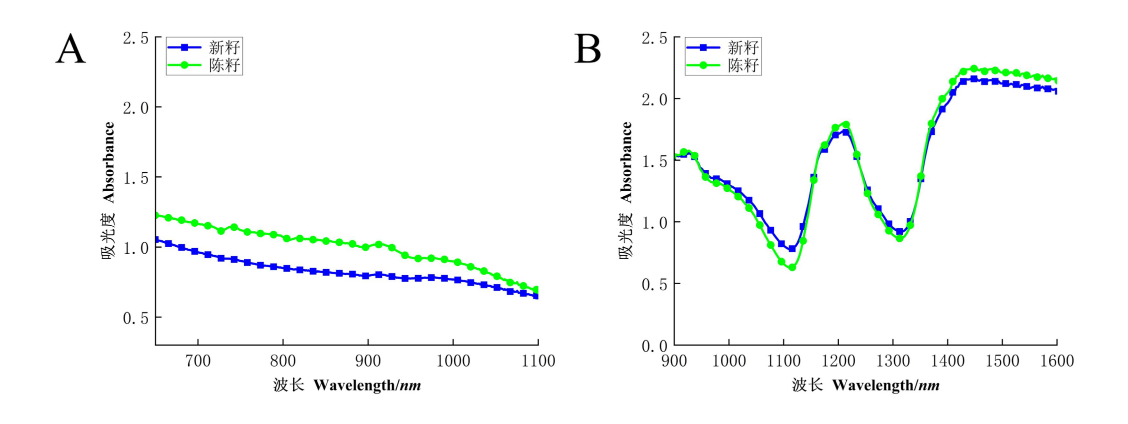


图 2 两种光谱仪采集的香榧新陈籽平均近红外光谱: (A) 光谱仪 1; (B) 光谱仪 2.

Fig.2 Average near-infrared spectra of fresh and stale Torreya grandis seeds collected by two spectrometers: (A) Spectrometer 1; (B) Spectrometer 2.

2.2 主成分分析

主成分分析主要作用是进行数据降维，以提取数据特征，通过提取的数据特征(主成分)可对两类样本间的差异进行初步分析。采用 PCA 方法提取两种光谱仪采集的香榧籽光谱的前 4 个主成分，方差贡献率和累计贡献率如表 1 所示。

表 1 香榧籽光谱的主成分贡献率

Table 1 Principal component contribution of Torreya grandis seeds spectra

	光谱仪 1 Spectrometer 1		光谱仪 2 Spectrometer 2	
主成分 Principal component	方差贡献率% Variance contribution rate%	累计贡献率% Cumulative contribution rate%	方差贡献率% Variance contribution rate%	累计贡献率% Cumulative contribution rate%
1	84.44%	84.44%	79.21%	79.21%
2	14.44%	98.89%	15.12%	94.33%
3	0.84%	99.73%	3.92%	98.25%
4	0.21%	99.94%	1.09%	99.34%

由表 1 可知，对于光谱仪 1，香榧光谱的前 2 个主成分的方差贡献率分别为 84.44％和 14.44%，累计贡献率达 98.89%。对于光谱仪 2，香榧籽光谱的前 2 个主成分的方差贡献率分别为 79.21％和 15.12%，累计贡献率达 94.33%。两类数据的前 2 个主成分的累计贡献率均达到 94%以上，说明 PCA 在降维和压缩特征的同时可以很好地保留原有光谱信息。此外，第 1 和第 2 主成分的方差贡献率远大于其他主成分，因此采用 PC-1 与 PC-2 为变量绘制散点分布图，以探究两光谱仪中香榧新籽和陈籽的光谱差异，其结果如图 3 所示。

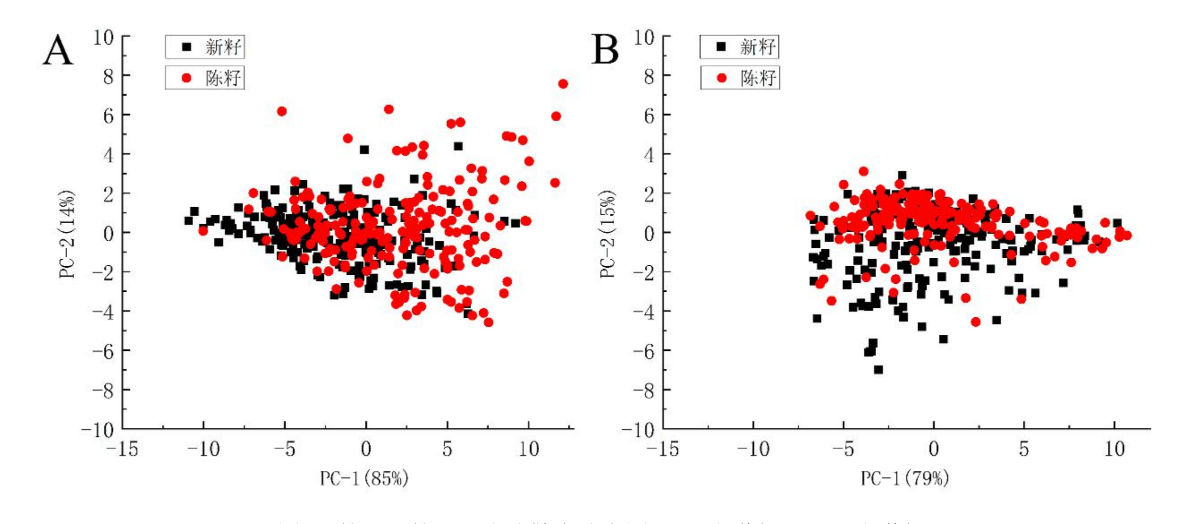


图 3 第 1 和第 2 主成分散点分布图: (A) 光谱仪 1; (B) 光谱仪 2.

Fig.3 Distribution of the first and second principal components: (A) Spectrometer 1; (B) Spectrometer 2.

由图 3 可以看出，对于光谱仪 1，香榧新籽的主成分散点在 PC-1 的 0 点右端有明显的簇拥现象，而香榧陈籽的主成分散点则分布更加分散；对于光谱仪 2，香榧新籽和陈籽的 PC-1 与 PC-2 主成分散点分布相互交

错，香榧陈籽的主成分散点更加聚集。此外，两光谱仪香榧新籽和陈籽的主成分散点均存在严重的重叠，未能很好地区分，需要采用化学计量学方法进行进一步分析。

2.3 预处理分析

采用 9 种预处理对两种光谱仪采集的香榧籽原始光谱进行预处理，然后应用 LDA 和 SVM 以及 BP 三类方法建立香榧新陈籽的判别模型，并利用预测集样本对判别模型精度进行验证，根据预测集的准确率以确定最优的预处理方法，其结果如图 4 所示。

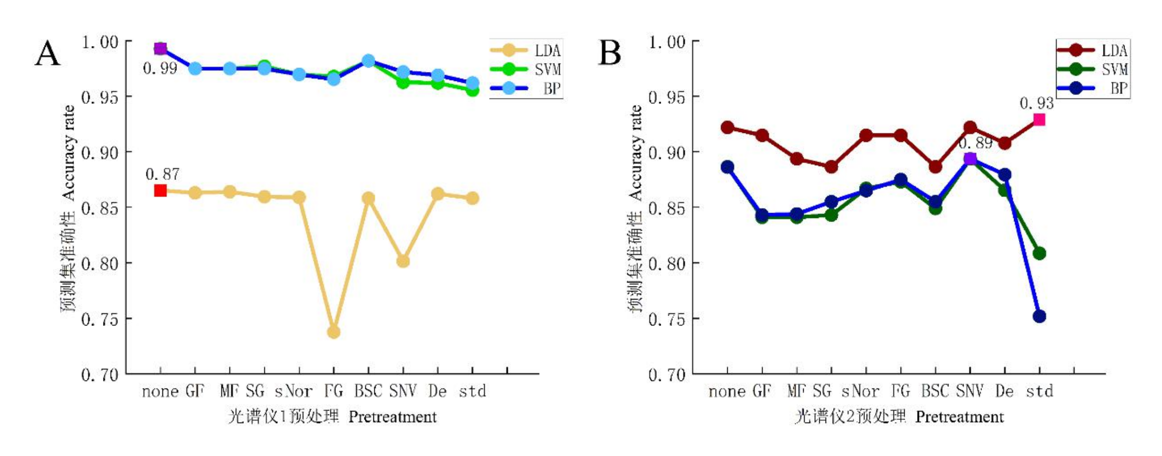


图 4 不同预处理下香榧陈籽判别模型的预测集准确率: (A) 光谱仪 1; (B) 光谱仪 2.

Fig.4 Accuracy of discriminative models for stale Torreya grandis seeds under different preprocessing methods: (A) Spectrometer 1; (B) Spectrometer 2.

由图 4(A)可知，对于光谱仪 1，经过预处理后，模型的性能大多仅能与原始光谱接近，均差于原始光谱所建立的模型性能。预处理后的所有判别模型性能均有不同程度地下降，尤其在 LDA 模型中，应用 FD（一阶导数）和 SNV（标准正态变换）方法导致模型效果显著下降。这可能是因为在预处理过程中，某些重要的特征峰被削弱或丢失，从而降低了 LDA 模型的判别能力。因此，对于光谱仪 1，原始光谱建立的香榧陈籽判别模型的性能最优。此外，在三种建模方法中，经预处理后，SVM 和 BP 神经网络模型性能变化趋势相似，但整体表现略有差异，其中 BP 模型性能略优于 SVM，而 LDA 模型性能则显著差于 BP 和 SVM 模型。

由图 4(B)可知，对于光谱仪 2，LDA 模型表现最优的预处理方法是标准化，而 SVM 和 BP 模型则表现最佳的预处理方法是 SNV。要注意的是，标准化仅对 LDA 模型适用并能显著提升其性能，而在 SVM 和 BP 模型中，标准化的效果较差。这可能是由于光谱数据的散射效应较小或基线漂移不显著，使用此类预处理方法可能会导致信号“过度修正”，从而削弱模型的表现。此外，SVM 和 BP 模型的预处理趋势相似，但 LDA 模型的效果在整体上优于这两种方法，且三种模型的预处理效果趋势基本一致。

表 2 为最优预处理方法下香榧陈籽的判别模型结果。由表 2 可知，对于光谱仪 1，BP 和 SVM 判别模型

的性能比 LDA 模型有明显的优势，BP 和 SVM 判别模型的预测集敏感性、特异性和准确性分别为 100%、98.25%和 99.29%。对于光谱仪 2，LDA 模型效果由于 SVM 和 BP 模型，其模型预测集敏感性、特异性和准确性为 90.12%、96.67%和 92.91%。此外，对比发现，两光谱仪数据的校正集和预测集结果均较为接近，表明建立的模型没有存在过拟合，有较好的稳健性。

表 2 最优预处理方法下香榧陈籽的判别模型结果

Table 2 Discriminative modelling results of stale Torreya grandis seeds under the optimal preprocessing method

数据集 Data set	光谱预处理 Spectral preprocessing	建模方法 Modeling methods	校正集 Calibration set			预测集 Prediction set		
			敏感性 Sensitivity	特异性 Specificity	准确率 Accuracy	敏感性 Sensitivity	特异性 Specificity	准确率 Accuracy
光谱仪 1	None	LDA	99.34%	97.67%	98.58%	73.13%	98.67%	86.62%
	None	SVM	99.22%	98.69%	98.93%	100.00%	98.25%	99.29%
	None	BP	98.44%	98.04%	98.22%	100.00%	98.25%	99.29%
光谱仪 2	standard	LDA	92.31%	91.89%	92.09%	90.12%	96.67%	92.91%
	SNV	SVM	89.12%	87.02%	88.13%	89.29%	88.24%	88.65%
	SNV	BP	88.36%	85.61%	87.05%	90.91%	88.37%	89.36%

2.4 特征波长选择

经最优预处理方法处理后，采用 CARS、ICO、SPA、VCPA 四种方法对于光谱仪 1 和光谱仪 2 的光谱数据进行特征波长选择，然后应用 LDA、SVM 和 BP 方法建立香榧陈籽的判别模型，并比较模型性能的优劣，以确定较优的特征波长选择。

2.4.1 CARS

图 5 为光谱仪 1 中香榧陈籽的 CARS 变量选择结果。图 5(A)为 CARS 变量选择过程中被选择的变量数量的变化。由图 5(A)可知，随着采样次数的增加，筛选出的波长变量数量逐级下降，0-5 次采样时下降趋势快，5-30 次采样时下降逐渐缓慢，30 次采样后趋于平稳。由图 5(B)可以看出，在 1-5 次采样过程中，RMSECV 值快速减少；在 6-23 次采样过程中，RMSECV 值缓慢下降，并在 23 次采样时其值最小；然后，随着采样次数的增加，RMSECV 值不断上升，表明此时有重要波长变量在采样过程中被剔除，从而导致模型性能下降。图 5(C)为 CARS 变量选择过程中波长变量的回归系数随采样次数增加的变化情况。图 5(C)中“*”所处的位置为 23 次采样，此时 RMSECV 值最小，即 23 次采样时所选择的波长变量集合最优。结合图 5(A)结果可知，最终共有 30 个波长变量被选择。图 6 为光谱仪 1 香榧陈籽的 CARS 标准差结果。图 6 直观地展示了每个样本中位值和标准差的平均贡献和稳定性，能够更容易地识别出哪些样本既具有较高贡献（高 MEAN 值）又具有较高稳定性（低 STD 值）。

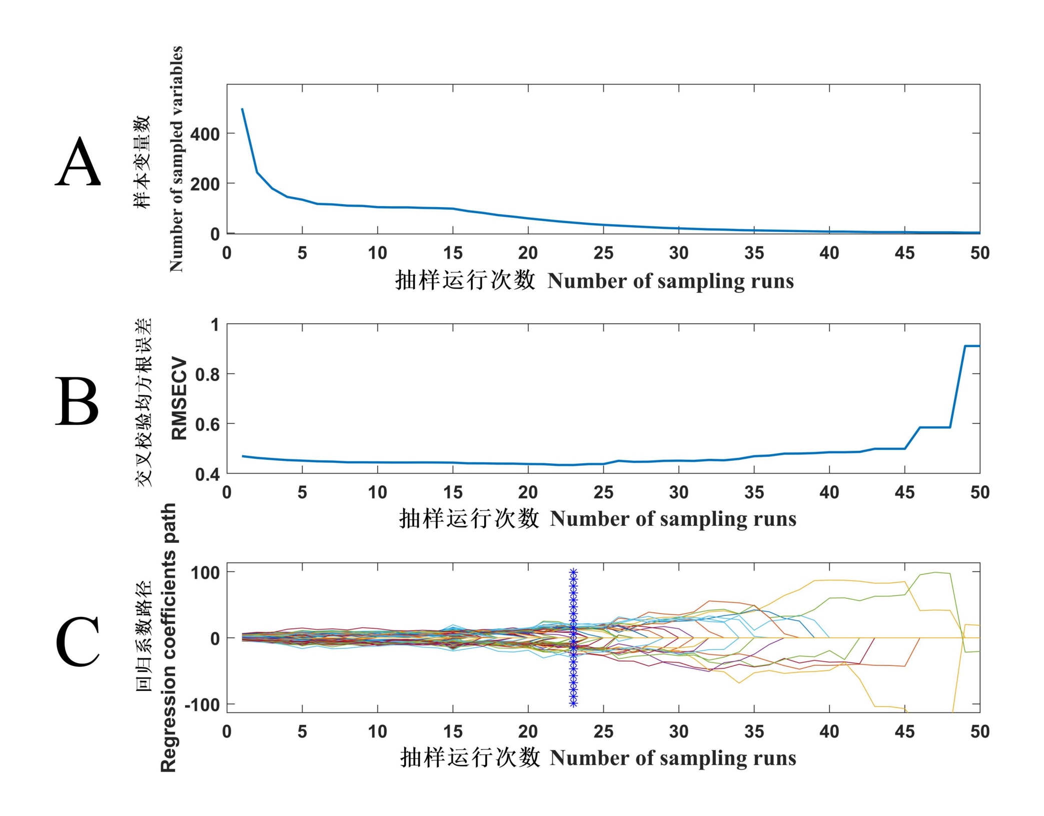


图 5 光谱仪 1 香榧陈籽的 CARS 变量选择结果:

(A) 被选择的波长变量数; (B) 交叉验证均方根差;(C)回归系数路径.

Fig.5 CARS variable selection results of stale Torreya Grandis seeds in spectrometer 1:

(A) Number of selected variables; (B) RMSECV; (C) Regression coefficient path.

光谱仪 2 中，对于 standard 及 SNV 预处理后的光谱，采用 CARS 方法分别选择了 32 个和 24 个波长变量。

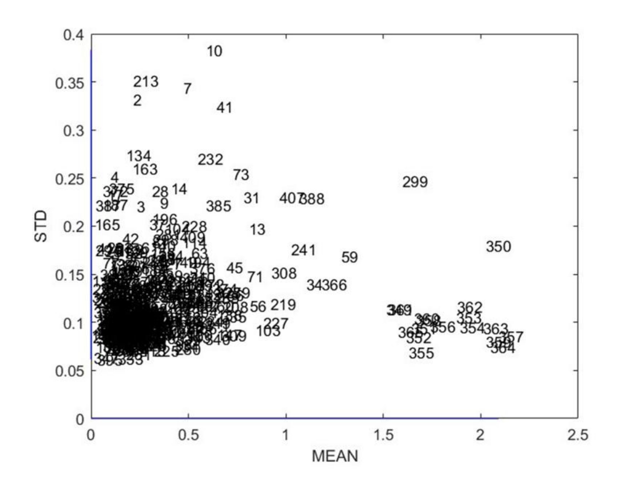


图 6 光谱仪 1 香榧陈籽的 CARS 标准差结果.

Fig.6 CARS standard deviation results of stale Torreya Grandis seeds in spectrometer 1.

2.4.2 ICO

图 7 为光谱仪 1 中香榧陈籽的 ICO 变量选择的结果。图 7(A)可知，在第 2 次迭代时 RMSECV 值最小，即此时的波段区间组合为最优，该区间组合共包含 172 个特征波长变量。图 7(C)为光谱仪 1 中香榧陈籽被选择的特征波段区间。由图 7(B)为 ICO 算法迭代过程中每个波长区间的采样权重随迭代次数增加的变化情况。图 7(B)中，颜色越黄则表示采样权重值越接近于 1；颜色越蓝则采样权重值越接近于 0；如果颜色介于蓝色和红色之间，则采样权重值处于 0 和 1 之间。图 7(A)为 ICO 方法迭代过程中 RMSECV 的变化情况。由图 7(C)可知，被选择的波段区间主要分布在 738-804nm 的长区间、873-896nm 和 920-904 两个小区间以及 986-1032nm 较长区间这四个区间范围。

光谱仪 2 中，对于 standard 及 SNV 预处理后的光谱，采用 ICO 方法筛选特征波段区间，最终分别选择了 119 个和 75 个特征波长变量。

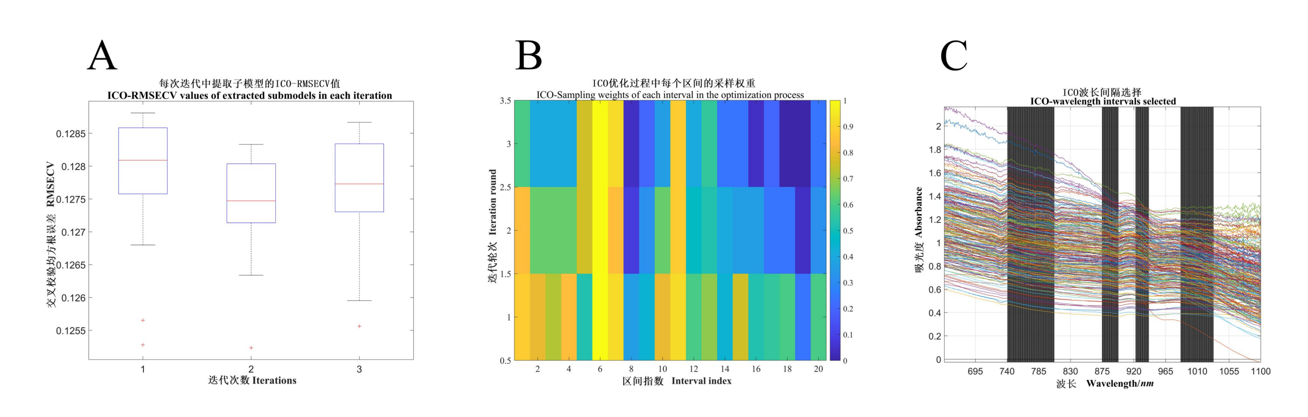


图 7 光谱仪 1 香榧陈籽的 ICO 变量选择结果:

(A) 交叉校验均方根误差; (B) 采样权重; (C) 区间变量选择.

Fig.7 ICO variable selection results of stale Torreya Grandis seeds in spectrometer 1: (A) RMSECV; (B) Sampling weight; (C) Interval variable selection.

2.4.3 SPA

图 8 为光谱仪 1 中香榧陈籽的 SPA 变量选择结果。SPA 以筛选波段少和筛选精度高而所熟知，可以看到 SPA 挑选的波长具有相当一部分的代表性。由图 8(A)中可知，当被选择的波长变量数由 1 增加到 4 时，模型的 RMSE 急速下降，表明所选择的波长均为香榧陈籽判别的重要波长变量。当被选择的波长变量数再继续增加时，模型的 RMSE 经历先下降再上升，而后再下降的过程，体现波长精选的过程；最终，当被选择的波长变量数为 14，模型的 RMSE 值最低。图 8(B)为 SPA 方法选择的 14 个波长变量的分布情况。由图 8(B)可知，所选择的波长主要分布在 1020-1100 nm 处的 7 个点，920 nm 波峰处靠近甲基 C-H 键三级倍频吸收峰，965 nm 为吸光度上升部分 938-975 nm 处挑选变量点，其他 5 个点均位于 650-875 nm 处吸光度下降部分。

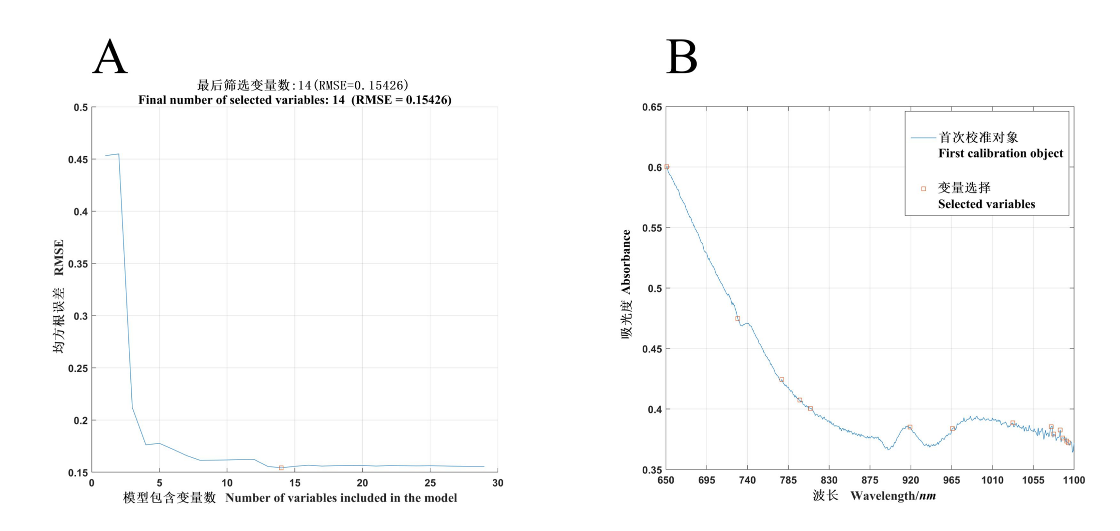


图 8 光谱仪 1 香榧陈籽的 SPA 变量选择结果：

(A) 均方根误差;(B) 变量选择分布.

Fig.8 SPA variable selection results of stale Torreya Grandis seeds in spectrometer 1:

(A) RMSE; (B) variable selection distribution.

光谱仪 2 中，对于 standard 及 SNV 预处理后的光谱，采用 SPA 方法分别筛选了 12 个和 17 个特征波长变量。

2.4.4 VCPA

图 9 为光谱仪 1 香榧陈籽的 VCPA 变量选择的均方根误差。由图 9 可知，在第 1-31 次迭代时，模型的均方根误差总体上呈快速下降趋势，表明一些干扰及无用信息波长变量正在被大量剔除。在第 31-37 迭代时，模型的均方根误差先上升而后下降，呈现波动趋势，表明正在进一步筛选变量。随着迭代的继续进行，模型的均方根误差逐渐上升，说明此时有部分重要变量被剔除，从而导致模型性能变差，均方根误差上升。在上述 50 次迭代中，第 34 次迭代时，均方根误差达到最小值，此时所选择的波长变量组合为最优，共有 14 个波长变量被选择。

光谱仪 2 中，对于 standard 及 SNV 预处理后的光谱，采用 VCPA 方法分别选择了 10 个和 12 个波长变量。

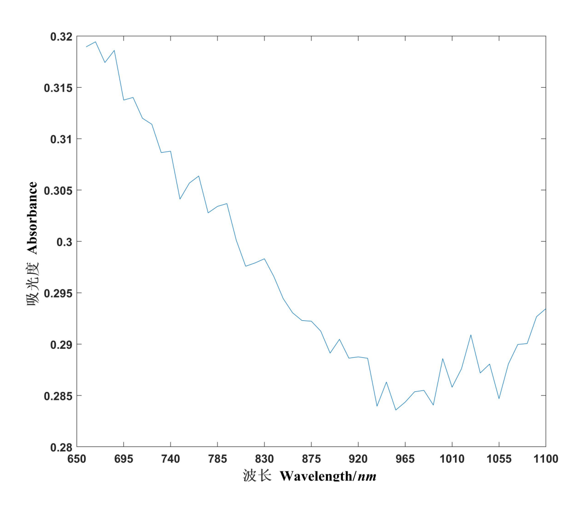


图 9 VCPA 均方根误差图.

Fig.9 VCPA root mean square error plot.

2.5 判别模型

根据 4 种变量选择方法筛选的特征波长，采用 LDA、SVM 及 BP 神经网络方法分别建立香榧陈籽的判别模型。表 3 为光谱仪 1 中基于特征波长的香榧陈籽的判别模型结果。由表 3 可知，与全波段模型相比，经 CARS 波长变量筛选后，三个模型的性能均略有提升，且所用的波长变量数仅为 30 个，占原波长变量数的 6%，有效简化了模型并提升了模型的稳定性；其中，CARS-SVM 和 CARS-BP 模型的性能较优，预测集的准确率分别为 100.00%和 99.29%。经 ICO 变量筛选后，模型的性能与全波段模型性能相当，但所用的波长变量数由 500 个减少为 172 个。经 SPA 及 VCPA 变量筛选后，部分模型的性能比全波段模型略有下降，但所用的波长变量数大大降低，均仅为 14 个。由此可知，对于光谱仪 1 的光谱数据，4 种变量选择均能有效筛选香榧陈籽的特征波长变量，其中 CARS 方法最优，ICO 方法次之，而 SPA 及 VCPA 则相对较差。此外，对比三种模型的性能，SVM 和 BP 模型的性能较优，预测集的准确率均在 98%以上，而 LDA 模型的性能则相对较差。综合上述分析，对于光谱仪 1，CARS-SVM 模型性能最优，预测集的敏感性、特异性及准确率分别为 100.00%、100.00%和 100.00%。

表 4 为光谱仪 2 中基于特征波长的香榧陈籽的判别模型结果。由表 4 可知，与全波段模型相比，经 CARS 变量筛选后，LDA 模型的性能有大幅下降，BP 模型的性能略有下降，而 SVM 模型的性能则有所提升，建模变量数由 218 个分别减少为 24 个、32 个和 32 个。经 ICO 及 SPA 变量筛选后，建模变量数大大降低，但模型性能没有改善，反而有所降低。经 VCPA 变量筛选后，LDA 模型性能有所降低，而 SVM 和 BP 模型性能则有所提升，模型的预测集准确率分别由 88.65%、89.36%提升到 94.33%、95.04%，且建模的变量数仅为 12 个，占原波长变量数的 5.5%。由此可知，对于光谱仪 2 的光谱数据，VCPA 为较优的变量选择方法，能有效筛选香榧陈籽的特征波长变量。此外，对比三种模型的性能，仍是 SVM 及 BP 模型的性能由于 LDA 模型。

综上分析，对于光谱仪 2 的光谱数据，VCPA-BP 为香榧陈籽判别的最佳模型，预测集的敏感性、特异性及准确率分别为 98.18%　93.02%、和 95.04%。

对比表 3 和表 4 结果可知，经特征波长筛选后，光谱仪 1 所建立的模型性能仍优于光谱仪 2，与原始光谱的结果一致，这可能是光谱仪 1 含有可见及近红外的光谱信息，更有利于实现香榧陈籽的判别。此外，从图 2 中也可看出，光谱仪 1 中香榧新籽和陈籽的平均光谱差异大，而光谱仪 1 中香榧新籽和陈籽的平均光谱差异较小，与模型的结果也相符合。

表 3 基于特征波长的香榧陈籽的判别模型结果(光谱仪 1)

Table 3 Discriminative model results of stale Torreya grandis seeds based on characteristic wavelengths (spectrometer 1)

波长选择方法 Wavelength selection methods	建模方法 Modeling methods	变量数目 Number of variables	预处理 Preprocessing	校正集 Calibration set			预测集 Prediction set		
				敏感性 Sensitivity	特异性 Specificity	准确率 Accuracy	敏感性 Sensitivity	特异性 Specificity	准确率 Accuracy
CARS	LDA	30	None	100.00%	99.22%	99.64%	75.68%	100.00%	87.23%
	SVM	30	None	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%
	BP	30	None	100.00%	100.00%	100.00%	100.00%	98.25%	99.29%
ICO	LDA	172	None	100.00%	100.00%	100.00%	75.68%	98.51%	86.52%
	SVM	172	None	100.00%	99.35%	99.64%	100.00%	98.25%	99.29%
	BP	172	None	100.00%	99.35%	99.64%	100.00%	98.25%	99.29%
SPA	LDA	14	None	99.34%	99.22%	99.29%	75.68%	98.51%	86.52%
	SVM	14	None	99.22%	99.34%	99.29%	100.00%	98.25%	99.29%
	BP	14	None	96.85%	96.10%	96.44%	97.65%	96.43%	97.16%
VCPA	LDA	14	None	100.00%	99.22%	99.64%	75.68%	98.51%	86.52%
	SVM	14	None	100.00%	99.35%	99.64%	100.00%	98.25%	99.29%
	BP	14	None	100.00%	100.00%	100.00%	98.82%	98.21%	98.58%

表 4 基于特征波长的香榧陈籽的判别模型结果（光谱仪 2）

Table 4 Discriminative model results of stale Torreya grandis seeds based on characteristic wavelengths (spectrometer 2)

波长选择方法 Wavelength selection methods	建模方法 Modeling methods	变量数目 Number of variables	预处理 Preprocessing	校正集 Calibration set			预测集 Prediction set		
				敏感性 Sensitivity	特异性 Specificity	准确率 Accuracy	敏感性 Sensitivity	特异性 Specificity	准确率 Accuracy
CARS	LDA	32	Standard	91.54%	89.86%	90.65%	75.31%	85.00%	79.43%
	SVM	24	SNV	93.20%	91.60%	92.45%	92%	95.00%	93.62%
	BP	24	SNV	86.00%	85.16%	85.61%	90.74%	87.36%	88.65%
ICO	LDA	119	Standard	90.00%	91.89%	91.01%	81.48%	81.67%	81.56%
	SVM	75	SNV	84.87%	84.92%	84.89%	87.27%	86.05%	86.52%

	BP	75	None	80.00%	85.84%	82.37%	79.37%	87.18%	83.69%
SPA	LDA	12	Standard	74.62%	80.41%	77.70%	70.37%	68.33%	69.50%
	SVM	17	SNV	91.97%	84.40%	88.13%	92.31%	86.52%	88.65%
	BP	17	SNV	89.33%	89.06%	89.21%	88.14%	90.24%	89.36%
VCPA	LDA	10	Standard	88.46%	88.51%	88.49%	82.72%	83.33%	82.98%
	SVM	12	SNV	92.81%	95.20%	93.88%	96.43%	92.94%	94.33%
	BP	12	SNV	97.92%	94.78%	96.40%	98.18%	93.02%	95.04%

3 结论

本研究利用近红外光谱技术结合特征波长选择对香榧陈籽进行快速无损判别。采用两光谱仪以透射方式采集香榧光谱，使用 9 种预处理方法进行光谱预处理，利用 CARS、ICO、SPA、VCPA 四种波长选择方法筛选香榧陈籽的特征波长，并应用 LDA、SVM 和 BP 方法建立香榧陈籽的判别模型。研究结果表明，对于光谱仪 1，预处理未能有效提升模型性能，CARS 方法为较优的特征波长选择方法，CARS-SVM 为香榧陈籽的最优判别模型，其预测集的敏感性、特异性和准确率均为 100%。对于光谱仪 2，standard 为 LDA 模型的较优处理方法，SNV 为 SVM 及 BP 模型的较优预处理方法，VCPA 为较优的特征波长选择方法，VCPA-BP 为香榧陈籽判别的最佳模型，预测集的敏感性、特异性及准确率分别为 98.18%、93.02%、和 95.04%。此外，光谱仪 1 的模型性能总体上优于光谱仪 2。本研究可为香榧陈籽的快速无损判别提供一种较新的检测方法，为香榧品质提供技术支持。

致谢

不知已经度过多少在宽阔学习平台上吸取新知的日夜。感谢学校和学院，感谢我最敬爱的老师和我最亲爱的同学，他们在我这两年半的研究生生活中给我的关心与支持。经过一年多的研究，在孙老师的指导下，我的论文得以顺利完成。在研究的过程中，得到了许多人的帮助，在此致以诚挚的谢意。

首先，我要感谢我的导师孙教授。他从论文的选题、开题、撰写到最后定稿，我都得到了卜老师的悉心指导和耐心帮助。孙老师结合研究方法论课程，多次为我修改论文与我反复讨论，使我的论文得以逐步完善。卜老师严谨的治学学风、一丝不苟的工作态度深深影响着我，激励我今后要认认真真、兢兢业业的工作。在此，我对孙老师表示由衷的感谢和真诚的敬意。

其次，感谢和我朝夕相处的同学们。他们在论文写过程中与我进行了有益的讨论，并在论文排版和校对过程中提供了热情的帮助。还有我的同门及师弟们，在组会的时候积极的参与我论文修改的讨论，给予我很多有用的建议，使我的论文得以逐步完善。

最后，还要感谢研究生期间所有教过我的老师，是他们引导我深入了解和掌握了相关领域的专业知使我的理论知识和实践能力都有所提高，为我的研究工作提供了基础。识

在这里 再次诚势感谢所有陪我走过研空生习的老师和同学们!

References

[1] WANG Rui, WANG Mei-Juan, TANG Fu-Bin, SONG Li-Li, LOU He-Qiang, NI Zhang-Lin, ZHONG Dong-Lian, MO Run-Hong. Journal of Nuclear Agricultural Sciences, 2021, 35 (11): 2578-2588.

王蕤, 王玫鹃, 汤富彬, 宋丽丽, 娄和强, 倪张林, 钟冬莲, 莫润宏. 核农学报, 2021, 35(11): 2578-2588.

[2] HE Ci-Ying, LOU He-Qiang, WU Jia-Sheng. Journal of Zhejiang A&F University, 2023, 40 (04): 714-722.

何慈颖, 娄和强, 吴家胜. 浙江农林大学学报, 2023, 40(04): 714-722.

[3] LI Hanlin, XIAO Nan, SUN Tong, HU Dong. Food and Bioprocess Technology, 2024, 17(12): 5221-5241.

李翰林, 肖男, 孙通, 胡栋. 食品与生物加工技术, 2024, 17(12): 5221-5241.

[4] CORTéS V., BLASCO J., ALEIXOS N., CUBERO S., TALENS P. Trends in Food Science & Technology, 2019, 85(138-148).

[5] QIAO Lu, MU Ying-Chun, LU Bing, TANG Xiuying. Food Reviews International, 2023, 39(3): 1628-1644.

乔璐, 穆迎春, 陆冰, 汤修映. 国际食品评论, 2023, 39(3): 1628-1644.

[6] SQUEO Giacomo, CRUZ Jordi, DE ANGELIS Davide, CAPONIO Francesco, AMIGO José M. Current Opinion in Food Science, 2024, 59(101203).

[7] VELEZ-SILVA Natasha L., DRENNEN James K., ANDERSON Carl A. International Journal of Pharmaceutics, 2024, 650(123699).

[8] REN Shu-Hui, JIA Yun-Fang. Spectrochimica Acta Part A: Molecular and Biomolecular Spectroscopy, 2023, 287(122080).

任树辉, 贾芸芳. 光谱化学学报 A:分子与生物分子光谱学, 2023, 287(122080).

[9] LOU Q-Jia, LEI Mei, WANG Yu, WANG Shao-Bin, GUO Guang-Hui, XIONG Wen-Cheng, JIANG Ying, JU Tie-Nan, ZHAO Xiao-Feng, COULON Frederic. Science of The Total Environment, 2024, 928(172264).

娄启嘉, 雷梅, 王宇, 王少斌, 郭广辉, 熊文成, 姜英, 鞠铁男, 赵晓峰, COULON Frederic. 总体环境科学, 2024, 928(172264).

[10] LIU Shi-Yu, WANG Shu-Tao, HU Chun-Hai, ZHAN Shu-Jie, KONG De-Ming, WANG Jun-Zhu. Spectrochimica Acta Part A: Molecular and Biomolecular Spectroscopy, 2022, 277(121261).

刘世玉, 王树涛, 胡春海, 詹树杰, 孔德明, 王军竹. 光谱化学学报 A:分子与生物分子光谱学, 2022, 277(121261).

[11] HU Jia-Qi, MA Xiao-Chen, LIU Ling-Ling, WU Yan-Wen, OUYANG Jie. Food Chemistry, 2017, 231(141-147).

胡家琦, 马晓晨, 刘玲玲, 吴彦文, 欧阳洁,. 食品化学, 2017, 231(141-147).

[12] LIU Jie Research on nondestructive detection method of chestnut quality based on near infrared spectroscopy technology ed. 2011

刘洁. 基于近红外光谱技术的板栗品质无损检测方法研究. ed. 2011.

[13] ZHANG Yan, XIE Yan-Li, SUN Shu-Min, MA Ming-Yang, LI Lin-Lin Journal of Henan University of Technology (NATURAL SCIENCE EDITION), 2014, 35 (02): 54-58

张严, 谢岩黎, 孙淑敏, 马明扬, 李琳琳. 河南工业大学学报(自然科学版), 2014, 35(02): 54-58.

[14] CANNEDDU G., JúNIOR L. C., DE ALMEIDA TEIXEIRA G. H. J Food Sci, 2016, 81(7): C1613-1621.

[15] MOSCETTI Roberto, MONARCA Danilo, CECCHINI Massimo, HAFF Ron P., CONTINI Marina, MASSANTINI Riccardo. Postharvest Biology and Technology, 2014, 93(83-90).

[16] CARVALHO Lívia C., MORAIS Camilo L. M., LIMA Kássio M. G., LEITE Gustavo W. P., OLIVEIRA Gabriele S., CASAGRANDE Izabela P., SANTOS NETO João P., TEIXEIRA Gustavo H. A. Food Analytical Methods, 2018, 11(7): 1857-1866.

[17] GHOSH Satyabrata, MISHRA Puneet, MOHAMAD Siti Nur Hidayah, DE SANTOS Rosario Martín, IGLESIAS Belén Diezma, ELORZA Pilar Barreiro. Biosystems Engineering, 2016, 151(178-186).

[18] HUANG Tian-Cheng, CAI Gui-Ming, LIU Hu-Bin, FENG Zhi-Yue, ZHAO Long-Lian, LI Jun-Hui. Spectroscopy Letters, 2022, 55(9): 607-617.

黄天成, 蔡桂明, 刘湖滨, 冯志跃, 赵龙莲, 李俊辉. 光谱学快报, 2022, 55(9): 607-617.

[19] KONG Ling-Fei, WU Cheng-Zhao, LI Han-Lin, YUAN Ming-An, SUN Tong. Journal of Food Composition and Analysis, 2024,

134(106560).

孔令飞, 吴成招, 李翰林, 园明安, 孙通. 食品成分与分析学报, 2024, 134(106560).

[20] ZHANG Wen-Xiang, PAN Liao, LU Li-Xin. Food Control, 2023, 147(109562).

张文祥，潘辽，陆立新. 食品控制学报, 2023, 147(109562).

[21] LI Ming-Xuan, SHI Ya-Bo, ZHANG Jiu-Ba, WAN Xin, FANG Jun, WU Yi, FU Rao, LI Yu, LI Lin, SU Lian-Lin, JI De, LU Tu-lin, BIAN Zhen-hua. Food Chemistry: X, 2023, 20(101022).

李明轩, 史亚波, 张九八, 万新, 方军, 吴仪, 傅饶, 李渔, 李琳, 苏连琳, 纪德, 陆图琳, 卞振华. 食品化学学报: X, 2023, 20(101022).

[22] TANG Rong-Nian, CHEN Xiao-Feng, LI Chuang. Applied spectroscopy, 2018, 72(5): 740-749.

唐荣年, 陈小锋, 李创, 应用光谱学, 2018, 72(5): 740-749

[23] JIANG Hui, XU Wei-Dong, CHEN Quan-Sheng. Spectrochimica Acta Part A: Molecular and Biomolecular Spectroscopy, 2019, 214(366-371).

姜辉, 徐卫东, 陈全胜. 光谱化学学报 A:分子与生物分子光谱学，2019214（366-371）, 2019, 214(366-371).

[24] LI Mei-Wen, XIA Li-Ye, WU Qing-Tao, WANG Lin, ZHU Jun-Long, ZHANG Ming-Chuan. Data Technologies and Applications, 2024, ahead-of-print(ahead-of-print).

李美文, 夏丽叶, 吴庆涛, 王林, 朱俊龙, 张明川. 数据技术与应用, 2024, 印刷前出版.

[25] ALMOUJAHED Muhammad Baraa, RANGARAJAN Aravind Krishnaswamy, WHETTON Rebecca L., VINCKE Damien, EYLENBOSCH Damien, VERMEULEN Philippe, MOUAZEN Abdul M. Chemometrics and Intelligent Laboratory Systems, 2024, 245(105050).

[26] SCHREUDER J., NIKNAFS S., WILLIAMS P., ROURA E., HOFFMAN L. C., COZZOLINO D. Spectrochimica Acta Part A: Molecular and Biomolecular Spectroscopy, 2024, 322(124716).

[27] ZHANG Xu, SUN Jia-Liang, LI Peng-Peng, ZENG Fan-Yi, WANG Hui-Hui. LWT, 2021, 152(112295).

张旭, 孙家良, 李鹏鹏, 曾繁毅, 王慧慧. LWT 食品科学与技术, 2021, 152(112295).

[28] MAO Shu-Can, ZHOU Jun-Peng, HAO Meng, DING An-Zi, LI Xin, WU Wen-Jin, QIAO Yu, WANG Lan, XIONG Guang-Quan, SHI Liu. Food Packaging and Shelf Life, 2023, 35(101025).

毛书灿, 周俊鹏, 郝蒙, 丁安梓, 李欣, 吴文锦, 乔宇, 王岚, 熊光全, 石柳. 食品包装与保质期, 2023, 35(101025).

[29] TAN Bao-Hua, YOU Wen-Hao, HUANG Cheng-Xu, XIAO Teng-Fei, TIAN Shi-Hao, LUO Li-Na, XIONG Nai-Xue. Electronics, 2022, 11(21): 3504.

谭宝华, 游文豪, 黄成旭, 肖腾飞, 田世豪, 罗丽娜, 熊乃雪. 电子学, 2022, 11(21): 3504

Rapid and non-destructive discrimination of stale seeds of Torreya grandis based on near-infrared spectroscopy and variable wavelength selection

FAN Zheng-Xin, ZAN Jia-Jun, DU Yuan[1], SUN Tong*

(College of Optomechanical Engineering, Zhejiang A & F University, Hangzhou 311304)

Abstract Objectives: Due to the oxidation of unsaturated fatty acids during storage, the taste and quality of Torreya grandis seeds declines. Unscrupulous merchants, seeking huge profits, blend Torreya stale seeds with fresh ones for sale, infringing upon consumers' interests. A fast and non-destructive identification method is needed . Methods: In this research, near-infrared spectroscopy was used to conduct rapid and non-destructive discrimination on stale Torreya grandis seeds. Spectra of shelled Torreya grandis seeds samples were collected in the wavelength ranges of 200-1160 nm and 900-1700 nm using two near-infrared spectrometers. Nine methods were employed to preprocess the spectral data. Then, four wavelength selection methods, namely interval optimization selection algorithm (ICO), competitive adaptive reweighted sampling (CARS), successive projections algorithm (SPA), and variable combination population analysis (VCPA), were utilized to screen the spectral characteristic variables of stale Torreya grandis seeds. Linear discriminant analysis (LDA), support vector machine (SVM), and backpropagation neural network (BP) methods were applied to establish discrimination models for stale Torreya grandis seeds. Results: The results indicate that for spectrometer 1, the CARS method is the optimal wavelength selection method, and the CARS-SVM model exhibits the best performance, with sensitivity, specificity, and accuracy all reaching 100% in the prediction set. For spectrometer 2, standardization and SNV are superior preprocessing methods. The VCPA variable selection method outperforms the other three methods, and the established optimal model is VCPA-BP, with the sensitivity, specificity, and accuracy of the model's prediction set being 98.18%, 93.02%, and 95.04%, respectively. Conclusions: Thus, it can be concluded that the discrimination models established based on the data from both spectrometers can effectively discriminate stale Torreya grandis seeds, and the overall performance of spectrometer 1 is superior to that of spectrometer 2. This study can provide a detection method for the rapid and non-destructive discrimination of stale Torreya grandis seeds, effectively guaranteeing the quality of Torreya grandis seeds.

Keywords: Near-infrared spectroscopy; Torreya grandis seeds; Torreya stale seeds; Wavelength variable selection;

Discrimination model

Received ****-**-**; accepted **-**-**

Special funds for basic scientific research business expenses of Zhejiang Provincial Colleges and universities (No. 2021td002) and key R&D in Zhejiang Province (No. 2020c02019).